Running head: DETECTING CHANGES IN EXTREMIST LINGUISTIC STYLE

Detecting Psychological Change through Mobilizing Interactions and Changes in Extremist Linguistic Style

Laura G. E. Smith[1]*, Laura Wakeford[2], Timothy F. Cribbin[3], Julie Barnett,[1] Wai Kai Hou[4]

[1]University of Bath, Department of Psychology.

[2]University of Abertay, Division of Psychology.

[3]Brunel University London, Department of Computer Science.

[4]Education University of Hong Kong, Department of Psychology, Centre for Psychosocial Health.

Paper accepted for publication in Computers in Human Behavior on 7th February 2020

Acknowledgments:

This work was supported by UKRI [grant number AH/N008006/1] and a British Academy mid-career fellowship to the first author.

Abstract

Social media interactions are popularly implicated in psychological changes like radicalization. However, there are currently no viable methods to assess whether social media interactions actually lead to such changes. The purpose of the current research was to develop a methodological paradigm that can assess such longitudinal change in individuals' social media posts. Using this method, we analyzed the longitudinal timelines of 110 Twitter users (40,053 tweets) who had expressed support for Daesh (also known as Islamic State, or ISIS) and we compared them to a baseline sample of twitter timelines (215,008 tweets by 109 users) to investigate the factors associated with within-person increases in conformity to the vernacular and linguistic style of tweets that supported violent extremism. We found that conformity to both extremist group vernacular and linguistic style increased over time, and with mobilizing online interactions. Thus, we show how to detect within-person changes over time in social media data and suggest why these changes occur, and in doing so, validate a methodological paradigm that can detect and predict within-person change in psychological group memberships through social media interactions.

Keywords: Social media, social interaction, social identity, linguistic style, radicalization, extremism

Detecting Psychological Change through Mobilizing Interactions and Changes in Extremist Linguistic Style

1. Introduction

> **"**Social interaction provides the site in which shared ideas about grievance and social change can become the foundations for collective action. That is to say, the airing and sharing of common perceptions of a grievance are group-making and can create a shared basis for taking action to rectify that injustice." (Smith, Blackwood, & Thomas, 2020, p. 21)

Social interaction can have profound effects on attitudes and behavior (Lewin, 1953). For example, just five minutes of face-to-face ingroup interaction about outgroups can increase peoples' intergroup prejudice, discrimination, and their confidence to undertake discriminatory intergroup actions (Moscovici & Zavalloni, 1969; Smith & Postmes, 2009, 2011). This is not just attitude polarization: through social interaction about shared grievances, people can develop shared *social identification* (psychological attachment to and cognitive self-definition as a member of a group; Tajfel & Turner, 1979) with a group that seeks social change (see McGarty, Thomas, Lala, Smith, & Bliuc, 2014; Smith, Gavin, & Sharp, 2015; Smith, Thomas, & McGarty, 2015). For social interaction to have this effect, the interactions need to mobilize people around shared grievances and the need for collective action to bring about social change (Smith et al., 2020; Smith et al., 2015; Thomas, McGarty, & Louis, 2014).

Mobilizing interactions are communications with others that express perceptions of the social context, including intergroup relations and outgroup behavior; and emotional reactions to these circumstances, which together capture experiences of social or political grievances. This is more than simply expressing negative emotions: communicating

perceptions of anger, harm, and threat as well as discussing collective actions is mobilizing because those communications express the perception that there is an illegitimate or unjust intergroup situation that needs to be changed via collective action (see Haidt, 2007; Graham, Haidt, & Nosek, 2009). In essence, communicating about these emotions, perceptions, and actions together expresses a socially grounded grievance; each element requires each of the other aspects to provide context and clarity. In contrast, communicating feelings of anger, or of perceptions of threat alone may not be politically or socially relevant. In essence, mobilizing interactions are more than the sum of their parts: they involve an interplay between the meaning of the communication of different aspects of grievances (emotions, cognitions, actions) and these must be understood together (see van Zomeren, Postmes, & Spears, 2008, for a meta-analysis of motivations underpinning collective action).

The act of communicating about these socially grounded grievances can mobilize people to take action (Smith et al., 2020; Smith, et al., 2015; Thomas, & McGarty, 2018). By receiving social validation on these perceptions from others during discussions, individuals' subjective perceptions can transform into as-if-objective, shared grievances. Thus, people's grievances become collectivized, and this increases their confidence in acting on them (Festinger, 1954; Smith & Postmes, 2011). Engaging in mobilizing interactions therefore represents a specific psychological pathway to political engagement and potentially, extremism (Smith et al., 2020; Thomas, et al., 2014). Indeed, extremism can be understood as a form of collective action because group members are seeking social or political changes (Alimi, 2006; Sageman, 2017). As Smith et al. (2020) argue, "if violent extremism is a form of collective action, a corollary of this is that psychologically engaging with groups (that is, a collective) is part of every journey to violent extremism. Such psychological engagement can be captured as social identification: psychological attachment to and cognitive self-definition as a member of a group (following Tajfel & Turner, 1979; see Doosje et al., 2016)" (p. 1).

The purpose of this article was to investigate whether engaging in mobilizing online interactions could increase social identification with a group that is premised upon a set of grievances and the perceived need for violent extremist action: Daesh[1] (also known as Islamic State, ISIS, or ISIL). This is timely and pertinent, as in the wake of the global increase in terror attacks (Europol, 2017), politicians have directed social media companies to take action to prevent people being radicalized on their platforms (Department for Digital, Culture, Media, & Sport, 2019; UK Home Affairs Committee, 2017). The underlying assumption made by these politicians was that, because social media offer opportunities for supporters and sympathizers of extremist groups to gather virtually and connect with likeminded others (Scrivens, Davies, & Frank, 2018), social media interactions may cause psychological change, such as an increase in social identification with a group that endorses violent extremism.

The availability of data via some social media platforms' application programming interfaces (APIs) makes it possible to test research questions about the impact of online interactions on the psychology of individuals over time. However, despite the availability of social media data to researchers, there have been to date no viable methods to use social media data to examine, at scale, whether the process of interacting using social media leads to within-person changes. Whilst machine learning techniques can be used to detect specific types of posts (De Smedt, De Pauw, & Ostaeyen, 2018; Kaati, Omer, Prucha, & Schrestha, 2015; Scrivens, et al., 2018) and online behavior associated with specific offline activities (Brynielsson, Horndahl, Johansson, Kaati, Mårtenson, & Svenson, 2013), there has been no research that applies machine learning to longitudinal public social media data to detect

[1] In using "Daesh", rather than "Islamic State" or other associated acronyms, we understand the group as a terrorist organization rather than a legitimate state (caliphate). In doing so, we adopt the same nomenclature as many Western governments (including Britain, France, and Australia). The name 'Daesh' is derived from al-Dawlah al-Islamīyah fī al-ʿIrāq wa-al-Shām — and is used as a pejorative acronym; sounding similar to the Arabic words Daes ("one who crushes something underfoot") and Dahes ("one who sows discord"). This name is often deemed preferable to "Islamic State" as it separates the religion of Islam from the terrorist group.

within-individual change in these types of posts. Therefore, in the current research we (a) developed a method that could use longitudinal social media data to model within-person changes over time, and (b), examined whether engaging in mobilizing social media interactions was associated with these changes.

To do this, first we leveraged research that suggests that social identification with a group could be detected through both *what* people write (the semantic content, including vernacular) and *how* they write it (their linguistic style; see Chung & Pennebaker, 2007; Pennebaker, 2011). Linguistic style is the pattern with which people use function words, which are a collection of non-semantic grammatical word categories. These categories include, for example, pronouns, prepositions, articles, and conjunctions. The way in which people use function words reflects their social psychological states and social relationships, and can predict their behavior (Campbell & Pennebaker, 2003; Chung & Pennebaker, 2007; Pennebaker, Mehl, & Niderhoffer, 2003; Scholand, Tauscznik, & Pennebaker, 2010; Tausczik & Pennebaker, 2009).

Discussion forum users who have a high degree of social identification with that community will also adopt a high level of forum-specific vernacular (Shrestha, Kaati, & Cohen, 2017). That is, they will learn and use ingroup jargon. This means that in terms of *what* people write, their ingroup vernacular is a useful marker of their group identity. Furthermore, when an individual identifies with a social group and that group identity is salient, they will conform to the normative linguistic style of that group (Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, & Potts, 2013): a specific pattern of function words. Social identification can therefore be expressed both through use of both ingroup vernacular and specific group-normative patterns of function words.

A consequence of this is that changes in an individual's use of ingroup vernacular and linguistic style towards the profile of a specific group could be an indicator that they are

developing social identification with that group (i.e., they will begin to adopt the vernacular and style of that group because it represents group normative behavior, see Turner, 1991). This extends work on communication accommodation and linguistic style matching that shows that when people engage in conversation (both face-to-face, Giles, Coupland, & Coupland, 1991, and via computers, Danescu-Niculescu-Mizil, Gamon, & Dumais, 2011) they change their linguistic style to become more similar to that of their partner (Gonzales, Hancock, & Pennebaker, 2010). Indeed, linguistic style matching is positively related to group cohesiveness (Gonzales et al., 2010). Thus, we assumed that, to the extent that an individual Twitter user's ingroup vernacular and linguistic style became more similar to that of prototypical ingroup posts, the more they identified with that group. Whilst someone may deliberately change their vernacular to evolve and respond to changing contextual conditions, it is more difficult to control use of function words (Pennebaker, 2011; Pennebaker & King, 1999). Similar to nonconscious mimicry, people produce function words non-consciously. Therefore, changes in patterns of function word use is potentially a more robust test of psychological change than changes in vernacular.

We also drew upon insights from theory that suggests that the mechanism by which individuals develop social identification over time with a group that holds a grievance is through mobilizing social interactions (Smith et al., 2020; 2015; Smith, McGarty, & Thomas, 2018). First, to validate this claim, we conducted a pilot study to establish that communicating about threat, harm, anger, and collection actions is mobilizing, in that it is positively associated with collective action intentions.

2. PILOT STUDY

2.1 Method

2.1.1 Context. The pilot study was conducted with a sample of Hong Kong residents during the violent protests of 2019. Massive protests began in June 2019 and lasted several

months. They followed the introduction of a controversial bill that would allow extradition of people in the Hong Kong Special Administration Region (SAR) to People's Republic of China. Despite the Hong Kong government's suspension and subsequent withdrawal of the bill, large-scale demonstrations and protests by Hong Kong residents rapidly transformed into regular pro-democracy demonstrations across districts in Hong Kong. Generally peaceful protests escalated to violent clashes among citizens, and between citizens and police, on the streets and in shopping malls, siege and occupation of governmental buildings and universities, and break-in and vandalism of pro-China shops and banks. Due to its massive scale, the unrest has been described as a potential public health crisis (Hou & Hall, 2019).

2.1.2 Design. This study had a simple, one factor design in which participants were randomly allocated to one of two communication conditions: mobilizing versus control. In both conditions, we asked participants to write 50-100 words, but the topic we asked them to write about varied across the conditions. In the control condition, we asked participants to, "describe your opinion of the current situation in Hong Kong". In the mobilizing communication condition, we asked participants to, "describe the anger you feel about the current situation in Hong Kong. In particular, please describe your perceptions of the harm and threat experienced by Hong Kong citizens". This was designed to elicit mobilizing content. We hypothesized that participants would write more content pertaining to harm, threat, anger, and collective action in the mobilizing (versus control) condition (H1), and that mobilizing content would be positively associated with collective action intentions (H2).

2.1.3 Participants. To establish the required sample size for a simple mediation model in which the communication condition predicted collective action intentions via mobilizing interactions, we conducted a power analysis using the continuously varying sample size approach to Monte Carlo power analysis with 5000 replications (Schoemann, Boulton, & Short, 2017). This suggested that approximately 151 individuals were required to

ensure statistical power was at least 80% for detecting the hypothesized indirect effect. We recruited 155 participants to allow for any participants that need to be excluded (e.g., for incomplete measures). Participants' mean age was 24.74 years (*SD* = 6.18). Thirty-three percent of the participants identified as male, 65% identified as female, and 2% indicated that they would "prefer not to say". In the five months prior to completing the survey, 72% of participants had attended a protest in Hong Kong. Sixty percent identified as a Hong Konger, 30% as Chinese, and 10% chose not to declare their national identity. Please note that "Hong Konger" is an internationally recognized identity but not nationality, therefore with this item we captured social self-identification as Hong Konger versus Chinese rather than an "official" nationality.

2.1.4 Measure of collective action intentions. To capture collective action intentions, we used a 5-item standardized scale (α= .86) adapted from Moskalenko and McCauley (2009). Items were, [I intend to] "encourage family and friends to sign a petition to stand up for the cause"; "write a letter to my local politician"; "join a peaceful protest"; "donate to an organization that fights for justice for Hong Kongers in a lawful way"; "donate to an organization that fights for justice for Hong Kongers but that sometimes breaks the law".

2.1.5 Mobilizing interactions. The variable for mobilizing interactions included communication content pertaining to harm, threat, anger, and collective action. To capture this content, we applied the standard LIWC2015 dictionary categories for anger and the moral foundations dictionary for harm (Graham, Haidt, & Nosek, 2009) to participants' written responses to the question about the political situation in Hong Kong. We used

validated custom dictionaries to detect content pertaining to threat and collective action[2] (Smith et al., 2018). Participants wrote 61 words on average (*SD* = 42.56), and the number of words they wrote did not differ significantly across conditions (control condition: *M* = 58.43, *SD* = 38.87; mobilizing condition: *M* = 63.69, *SD* = 46.27), $F(1, 153) = 0.59$, $p = .44$.

Table 1. Factor Loadings and Descriptive Statistics for Mobilizing Interactions Variable (*N* = 155)

Observed variable (dictionary categories)	Factor Loading	*M*	*SD*	1.	2.	3.	4.
1. Harm	0.89	2.05	2.81	-			
2. Threat	0.21	0.43	1.17	.19*	-		
3. Anger	0.61	3.67	4.30	.54**	.13	-	
4. Collective Action	0.18	2.16	2.36	.16*	-.04	.13	-

Note. *$p<.05$, **$p<.001$.

Then, we performed an exploratory factor analysis (EFA) with principal axis factoring and oblimin rotation on the observed LIWC variables (Table 1) using the Psych and GPARotation packages in *R*. The EFA indicated that a single factor solution was the most appropriate for the data (with one factor's eigenvalue > 1 and all other eigenvalues < 0.50). This single factor accounted for 31% of the variance in the harm, threat, anger, and collective action variables and had a sum of squared loading of 1.24. The root mean square of the

[2] There were 328 words in total in the four dictionaries, of which 14 words overlapped across dictionaries: *abuse**, *attack**, *brutal**, *cruel**, *destroy**, and *kill* were in anger and harm; *endanger** and *harm** were in harm and threat; *fight* was in collective action and harm; *hostil** and *threat** were in anger and threat; *protest* was in anger and collective action; *war*, *wars*, and *warring* were in harm and anger. This overlap did not substantively alter the results.

residuals (RMSR) was 0.03; Tucker Lewis Index of factoring reliability = 1.04. We then used the factor scores as a predictor variable that we called, "mobilizing interactions". The factor scores for this latent construct represent and contextualize the expression of anger, harm, and threat as collective grievances.

2.2 Results and Discussion

To test for causal mediation, whereby communication condition predicted collective action intentions through increases in mobilizing interactions[3] (controlling for prior protest attendance and self-reported identity as a Hong Konger vs. Chinese), we ran Hayes' PROCESS 3.4 macro (model 4) with 5000 bootstrap samples (Hayes, 2018). The model explained a significant amount of variance in collective action intentions, $R^2 = .38$, $p < .001$. There was a significant positive direct relationship between communication condition and mobilizing interactions, whereby participants communicated more mobilizing content in the mobilizing condition ($M = 0.19$, $SD = 1.11$) versus control condition ($M = -0.18$, $SD = 0.63$), $\beta = 0.37$, $p = .02$ (95% Bias-Corrected Bootstrapped Confidence Intervals or *CI*s = 0.07, 0.68). This suggested that our manipulation was successful and provided support for H1. The indirect effect of communication condition on collective action intentions through mobilizing interactions was positive and significant ($\beta = .08$; 95% *CI*s = 0.01, 0.16), supporting H2. The direct effect of condition on collective action intentions was non-significant in the presence of the mediator ($\beta = 13$; 95% *CI*s = -0.20, 0.47; $p = .43$). These results suggested that communicating about harm, threat, anger, and collective action was mobilizing, in that it increased support for undertaking collective action.

[3] To further verify that mobilizing interactions operate as a single construct, we entered the observed variables anger, harm, collective action, and threat separately (instead of the mobilizing interactions variable) at Step 2 of a regression model predicting collective action intentions (where condition was entered at Step 1). None of these individual variables were significantly related to collective action intentions. However, when the factor scores for the mobilizing interaction construct were entered into the model instead of the four observed variables, it was significantly related to collective action intentions ($\beta = 0.14$, $p = .05$) and significantly contributed to the variance explained, R^2ch = .02, Fch(1,145) = 4.04, $p = .05$.

3. MAIN STUDY

3.1 Method

3.1.1 The Context

In this main study, we investigated whether engaging in mobilizing online interactions over time facilitated individuals' socialization as supporters of an extremist group: Daesh. We hypothesized that individual Twitter users would conform more to the vernacular and linguistic style of tweets by supporters of Daesh that endorsed violent extremism the more they engaged in mobilizing interactions on Twitter.

We collected a corpus of tweets by individual user accounts that showed support for Daesh. In 2014 – 2015, Daesh had reached its peak penetration on Twitter, with an estimated 46,000 – 90,000 Daesh supporters on the platform (Stern & Berger, 2015)[4]. On Twitter, Daesh supporters shared links to extremist material that was hosted by other sites and engaged in opinion dissemination and discussion. At the time of data collection for the current project (2016/7), Twitter was established as a gateway platform identified by Sunni extremists as an important platform for their missionary work. For example, the Shumukh al-Islam forum released a "Twitter Guide", in which they outlined why their missionaries should use Twitter to recruit new supporters (see Ashcroft, Fisher, Kaati, Omer, & Prucha, 2015). Thus, members of Daesh used Twitter to recruit new supporters. Therefore, by collecting longitudinal Twitter data, we would be able to examine the mechanisms of interaction on Twitter by which supporters of Daesh were recruited and radicalized. This made Twitter an appropriate platform on which to test our method and hypothesis.

[4] These figures should be understood in relation to the nature of the feature set (online behaviors) that was used to identify those Twitter accounts as supporting Daesh. These feature sets vary across different research projects, with some researchers defining Daesh supporters not by the specific kinds of semantic content they post, but by who their followers are (see Berger & Morgan, 2015).

3.1.2 Data Collection

3.1.2.1 Feature Validation. To ensure that we identified a temporally valid sample of English-speaking[5] Twitter users who explicitly expressed support for Daesh, we systematically identified and validated the features (online behaviors) that were associated with expressing support for Daesh on Twitter at the time of the research, via a systematic four-phase process. The first phase was a literature review of articles that identified social media features associated with support for Daesh. The search terms "Islamic State" and "social media" yielded 26 articles (across SCOPUS, Web of Science, and PsychINFO). The second phase was consultation with terrorism and extremism experts from government, and the security and defence sectors. The combination of Phase 1 and 2 provided an initial set of 65 features associated with expressing support for Daesh. In Phase 3, we manually inspected examples of these features on the live Twitter stream to confirm whether or not they could be used to manually identify Twitter users who claimed to support Daesh. The fourth phase was consultation with members of Muslim community to validate that the features identified reliably discriminated between the online behavior of Daesh supporters and mainstream Twitter users. We excluded the features that as a result of the consultation process were not deemed to be discriminatory. This led to a final set of 30 features, 22 of which were manually observable by a human rater at the individual Twitter user level in the live Twitter stream (specific details of the features are not described here for ethical reasons[6]). Manually observable features included (but were not restricted to): retweeting Daesh propaganda,

[5] Our aim in targeting English-speaking Twitter users was to develop an algorithm that was appropriate for this population and to provide proof of concept for the method, rather than to create an algorithm that generalized across Daesh supporter populations who communicated in other languages.

[6] We have chosen not to publicly disclose the feature set or the Daesh vernacular dictionary due to the potential that they could be used incorrectly to falsely identify individuals as supporters of violent extremism. The accuracy of the features reported here is likely to change over time. To establish the temporal stability of the features and their predictive validity, future research should use longitudinal ground truth data (independent verification of the identity of the users) to validate them. The purpose of this research was not to create a time-independent, validated algorithm or lexicon but to develop a new method to explore changes in linguistic style over time.

celebrating suicide bombers, displaying the Daesh flag or rejoicing following a Daesh terror attack. These features could include profile pictures/avatars or may have been within the profile/Twitter account description.

Using Chorus Tweetcatcher 1.3 software (Brooker, Barnett, & Cribbin, 2016), we searched Twitter for tweets that included the tweet-level manually-observable features. These tweets were then inspected to determine whether (or not) they displayed Daesh-supporting features; if they did, their associated user accounts were directly inspected both via the Twitter website and by extracting their publicly available timelines (tweets and associated metrics) from Twitter's API using Chorus software. Features that could be detected via Chorus included specific hashtags and topics of discussion (such as martyrdom operations, or the death of a senior Daesh figure) that may not have been visible on a user's feed on the Twitter website (which only shows relatively recent posts), but were available once user timelines were retrieved via the API.

Chorus uses the Twitter API to harvest (up to) the most recent 3200 tweets of individual users. So, whilst they may have expressed support for Daesh at the time of data collection, 3,200 tweets back in time (weeks or possibly months earlier) they may not have been as extremist and/or their focus and narrative may have been different. In fact, if the assumption is correct that people and narratives can change (e.g., become more extremist) through increased online engagement, then their tweets should be less extremist at the start of their timeline, and more extremist in more recent tweets. The search yield for any user is limited by the number of tweets that are publicly available for that user name (therefore, recently opened accounts would normally yield fewer posts).

User profiles and live timelines were manually inspected to determine whether (or not) they also displayed Daesh-supporting features. The coding was binary: user accounts were either coded as expressing support for Daesh (and included in the dataset) or not. We

determined the coding criterion for Daesh support as a minimum of two features through the validation process described in 2.2.1 above. Using co-occurrence of features (rather than single features) ensured that all features were understood not in isolation but in the context of their co-occurrence with other features. This also helped to alleviate ethical concerns about falsely identifying an account as Daesh-supporting (someone might retweet a Daesh-sympathetic post once but this might not be as meaningful as if they also use the Daesh flag as their avatar). Through this procedure, a human rater identified 110 Twitter accounts that explicitly expressed support for Daesh. A second coder independently coded the 110 accounts, and inter-rater reliability was 100%. We do not make any claims or assumptions about the users in this sample beyond the fact that each user expressed support for Daesh in their tweets and/or profiles. Since concluding data collection, all of the 110 Daesh-supporting user accounts have been suspended by Twitter for violating the terms of use of the platform.

We also retrieved the longitudinal Twitter data of 109 users to provide a baseline against which to compare the behavior of the Daesh supporter sample. This provided a non-extremist baseline of the features that would be common to both mainstream Muslims and Daesh supporters, such as English transliterations of Arabic and content pertaining to Islam. Therefore, we retrieved the longitudinal data of 109 Twitter users who had used the hashtag #Ramadan. Through using this as the baseline sample/class, we endeavored to ensure that features associated with mainstream Islam were not conflated with those that indicated support for Daesh[7]. Discriminatory ability of the features was verified via 2 x 2 chi-square tests that found significant differences in the co-occurrence of specific features across the two samples.

[7] The baseline sample was thus a control group only insofar as it provided a baseline for these features. We could compare the Daesh-supporter sample to a variety of different groups, and each time our algorithm would produce different coefficients. However, the absolute value of the coefficients is not the outcome of interest: it is the change in Daesh-supporter tweet classification relative to the baseline over time that is key to indexing change via our method.

3.1.2.2 Sample. The data collection strategy resulted in a total sample of 255,061 tweets in English (40,053 tweets by 110 Daesh supporters, and 215,008 tweets by 109 users in the baseline sample; 16 tweets were omitted due to missing data; the data for the analyses described below are available in the Mendeley Database, doi:10.17632/jz9hm3n49d.1). An a priori power analysis established that the sample size required for a medium effect size (F^2 = 0.15) was 189 observations (to detect significance at $p < .05$ in a linear regression model with 13 predictors). The Daesh-supporter tweet sample was smaller because the Daesh-supporting accounts were "younger" with fewer tweets than the baseline accounts. This is likely to be a result of Twitter's effort at the time to remove Daesh-related activity. A high proportion of Daesh-supporters had multiple, short-lived sequential accounts, because previous incarnations of the account were suspended by Twitter. We control for multiple accounts by the same user in our analyses.

3.1.3 Natural Language Processing of Tweet Content

3.1.3.1 Function words. Following linguistic style matching research (see Gonzales et al., 2010), we used Linguistic Inquiry and Word Count (LIWC2015) software (Pennebaker, Booth, & Francis, 2007) to establish the proportion of words within each tweet that appeared in the LIWC function words dictionary categories (lexicons). The LIWC function word categories each contain words that cluster together according to their grammatical function (e.g., pronouns, articles, prepositions; Biber, 1988; Pennebaker, Boyd, Jordan, & Blackburn, 2015; Tausczik & Pennebaker, 2009). There was no overlap between the content of these lexicons and the features that we used to identify the sample of Daesh-supporters' Twitter accounts. The function word categories are described in Table 2.

The pattern of function words that a group member uses indicates their knowledge of a specific social psychological context. This is because to use and understand certain function words, the interlocutors must have a shared "insider" knowledge of spatial and temporal

relations, intergroup relations, the salient outgroup(s), and so on. For example, pronouns can indicate whether an individual's focus is on their personal identity (first-person singular

Table 2

Meaning of LIWC Function Word Categories

LIWC Category and examples	Meaning
Prepositions ("to", "with", "above")	Prepositions are referential and help describe spatial and temporal relations (Lucic & Bridges, 2018).
Third person plural pronouns ("they", "their", "they'd")	Third person pronouns are markers to suggest that the speaker is aware of a specific outgroup.
Articles ("a", "an", "the")	Articles are referential. To know the meaning of an, the, etc. demands that the speaker and listener have a shared understanding of the specificity of the context (the use of "a" versus "the").
Auxiliary verbs ("is", "do", "have")	Auxiliary verbs are part of a passive (versus active) way of speaking; signal lower power (Pennebaker, 2011).
Second person singular pronouns ("you", "your", "thou")	Second person pronouns are markers to suggest that the speaker is socially engaged or aware.
First person singular pronouns ("I", "me", "mine")	Use of first person singular is associated with age, sex, depression, illness, and more broadly, self-focus; self-attention; honesty.

First person plural pronouns (“we”, “us”, “our”)	First person plural can be a marker of group identity/community (Pennebaker & Lay 2002).
Negations (“no”, “not”, “never”)	Negations provide assertions with specific polarity (truthfulness or falseness). Negations are easier to understand, and to formulate, when interlocutors have a shared understanding of the possibilities to be negated (Khemlani, Orenes, & Johnson-Laird, 2012).
Conjunctions (“and”, “but”, “whereas”)	Conjunctions are operative terms that connect and qualify two sentence components to provide a conditional, truth-functional meaning (Bott, Frisson, & Murphy, 2009).
Impersonal pronouns (“it”, “its”, “those”)	Refer to the object of discussion. To know the meaning of it, its, those, etc. demands that the speaker and listener have a shared understanding of the object of the conversation.
Quantifiers (“few”, “many”, “much”)	Convey specific details of the context and help describe scale/quantities.
Adverbs (“very”, “really”)	Adverbs modulate the meaning of verbs, adjectives, other adverbs, and noun phrases and thus evaluate what is being spoken of (Quirk, Greenbaum, Leech, & Svartvik, 1985).

pronouns such as “I”, “me”, “my”), or on their identity as a group member (first-person plural pronouns such as “we”, “our”). Use of third person plural pronouns (“they”, “them”) requires interlocutors to have shared knowledge of who the relevant outgroups are (Lyons, Aksayli, & Brewer, 2018). Prepositions describe spatial and temporal relations (Lucic & Bridges, 2018): To know the meaning of over, on, to, etc., requires that the interlocutors understand the

relative, real, or symbolic location of the subject of the communication (Pennebaker, Mehl, & Niederhoffer, 2003). Whilst each category of function words has unique psychological meaning, it is the pattern of them in relation to each other in a text that represents the ingroup norm for linguistic style (see Gonzales, Hancock, & Pennebaker, 2009) and ingroup context and its central players. Therefore, we were interested in how much users adhered to a particular pattern of function words, or a particular linguistic style, overall, and how this changed over time (rather than the raw scores of or correlations between each category of function words, *per se*).

3.1.3.2 Mobilizing interactions. To capture mobilizing interactions within tweets by Daesh supporters[8], we used the same method as in the pilot study. First, we performed an EFA with principal axis factoring and oblimin rotation on the observed LIWC variables, harm, threat, anger, collective action, and Twitter interactions (@mentions; Table 3). The EFA indicated that only one factor had an eigenvalue above 1.00 and this factor accounted for 29% of the variance in the harm, threat, anger, collective action, and Twitter interactions variables, and had a sum of squared loading of 1.45. The root mean square of the residuals (*RMSR*) was 0.03, the Tucker Lewis Index of factoring reliability was 0.93; RMSEA = 0.06. We then saved the factor scores for the single factor to use as a predictor variable named mobilizing interactions.

To establish face validity of the mobilizing interactions construct, we manually coded a subset of tweets (Table 4) for whether or not they were mobilizing. The factor scores for mobilizing content ranged from -0.58 to 16.60, with a median of -0.49. We selected 25 tweets

[8] A factor analysis on the baseline sample indicated that the latent variable was not a good fit for those data: the factor loading for twitter interactions (@mentions) was 0.01 (compared to 0.07 in the Daesh-supporters sample). This implies that users in the baseline sample engaged in very little interaction with other users on Twitter about harm, threat, anger, and collective action. This is unsurprising given that (unlike the Daesh-supporter sample) the baseline sample of tweets was not authored by a psychological group with a grievance around which to interact and mobilize. Therefore, there was no conceptual or statistical justification for computing the mobilizing interactions variable for the baseline tweet sample.

randomly from each quartile of factor scores, and two coders independently coded each tweet, with a third coder resolving any disagreements. We used two coding categories: a) mobilizing; b) not mobilizing. The tweets were coded as mobilizing (versus not mobilizing) if they included any content pertaining to harm or threat to Muslims, anger, and collective action. We computed Cohen's kappa (Cohen, 1960) for the coder pair to provide an index of inter-rater reliability for each quartile. The resulting kappas indicated an acceptable level of inter-rater agreement (Table 4; Landis & Koch, 1977).

Table 3

Factor Loadings and Descriptive Statistics for Mobilizing Interactions Variable (Daesh Supporters Test Sample, *n* = 38,053)

	Observed variable (dictionary categories)	Factor Loading	*M*	*SD*	1.	2.	3.	4.	5.
1.	Harm	0.62	0.21	1.36	-				
2.	Threat	0.12	0.02	0.41	0.01**	-			
3.	Anger	1.00	0.54	2.36	0.64***	0.17***	-		
4.	Collective Action	0.17	0.38	1.76	0.09***	0.01	0.18***	-	
5.	Twitter Interactions	0.07	0.77	0.42	0.06***	0.00	0.05***	0.01	-

Note. **p = .01, ***p<.001

As expected, in the fourth quartile, which containing tweets with the highest factor

scores, there was a significantly greater number of tweets coded as mobilizing than not mobilizing, χ^2 (1) = 4.84, p = .03. In the third and second quartiles, there was no difference between the number of tweets coded as mobilizing and not mobilizing, (Q3: χ^2 (1) = 2.00 p = .35; Q2: χ^2 (1) = 2.00, p = .35). In the first quartile, containing the lowest factor scores, there was a significantly greater number of tweets coded as not mobilizing than mobilizing, χ^2 (1) = 14.44, p < .001. This suggested the variable had appropriate face validity.

Table 4

Coding of Tweets for Mobilizing Interactions

		%			
Code		Q1	Q2	Q3	Q4
1	Mobilizing	12	40	60	72
0	Non-mobilizing	88	60	40	28
	κ	.78	.64	.76	.81

Note. Q = quartile of factor scores, whereby Q4 contained the highest factor scores (tweets containing the highest proportion of mobilizing content) and Q1 contained the lowest factor scores. κ = Cohen's Kappa.

3.1.3.3 Daesh vernacular. Through the literature review, and consultation process with terrorism and extremism experts from government and the security and defence sectors (as described in 3.1.2.1 above), we developed a new custom dictionary for Daesh vernacular (contents not disclosed for ethical reasons) that included derogatory terms aimed at non-Daesh supporters, Daesh-relevant terms, and transliterations of Daesh-relevant Arabic words. Taken together the combination of and overall prevalence of these words (rather than the use of each word individually) was associated with sympathy for Daesh. We validated the Daesh vernacular dictionary using an independent sample of 2,359 English language tweets that

included the hashtag #Dabiq (the Daesh official magazine), which were collected independently of this study in mid-2016. The mean percentage of words in the tweets in this dataset that were in the Daesh vernacular dictionary was 10.19 (*SD* = 6.68), the median was 10.00, with a range of 50. This suggests that the content of this dictionary generalized across relevant samples.

3.1.4 Analytic Strategy

Our analysis then proceeded via two steps. The first step was to train a learning algorithm on a gold-standard subset of tweets to discriminate between extremist tweets and non-extremist tweets. The application of this algorithm to the rest of the data corpus then provided each tweet with a probability that it was extremist versus not extremist. We assumed that within the sample of 40,053 tweets authored by Daesh users there would be variation in the extent to which those tweets would be extremist (defined as showing support for violent extremism, whether directly authored (in an original tweet) or not (retweeted)). To this end, we sought to exploit within-person variation in tweet-level linguistic style to model the factors that were positively associated with within-person increases in extremist linguistic style (as a proxy for psychological change). The second step therefore was to assess the extent to which users' tweets became more extremist over time (indexed via the probability of classification as extremist in step 1), and to identify the factors associated with such changes.

Therefore, for the first step we manually-coded a gold-standard training set of 2,000 extremist human-rated tweets from the Daesh sample and 2,000 non-extremist tweets from the baseline sample (as per Burnap, Gibson, Sloan, Southern, & Williams, 2016). The 2,000 extremist tweets for the training sample met the following criterion: The tweets endorsed violent extremism and/or Daesh (as a violent extremist group)[9]. All of these tweets included

[9] In setting this criterion, we adopted as a starting point the United States government's definition of violent extremists as "individuals who support or commit ideologically-motivated violence to further political goals,"

one or more of the following features: support for violent extremism, including support for Jihād (struggle/war on the path of God), Mujāhidīn (Jihadist fighters), and Istishhad (a martyrdom operation); support for Daesh, including al-walā' wa al-barā' (unconditional adherence to the Islamist project and the group supporting it, and disavowal and disassociation of other causes and groups supporting them), sharing Daesh propaganda (including "shoutouts" and Daesh Telegram channels), praise for the Caliphate (Khilafah) and criticism of Kufr (unbelief, disbelief) and Kuffār (unbelievers). Coding allowed for variations of spellings of these features. Whilst there was variation in exactly how extremist those features were, they were all confirmed as associated with support for violent extremism in our earlier four-step validation process, justifying a binary coding framework (extremist versus not extremist). Thus, all tweets in the training sample met the criterion. A second coder manually verified the coding of the training sample. We also randomly selected 2,000 tweets from the baseline sample that did not contain any of those extremist features. A coder manually verified that these baseline tweets were not extremist. This provided a total training sample of 4,000 gold-standard tweets for the classifier. The remaining dataset of 251,061 tweets formed the test sample.

All Daesh users both authored tweets directly and retweeted – in other words, it was normative Daesh-supporter behavior to both compose original tweets and retweet (see Supplemental Materials). For this reason, our training sample included extremist tweets that were both directly authored (52.3%, n = 1046) and retweeted (47.7%, n = 954). The ratio of directly authored tweets to retweets in the training sample reflected the fact that directly authored tweets were on average more extremist than retweets and therefore, there were more directly authored tweets that were suitable to include in the training sample. The same

(Executive Office of the President of the United States, 2016). We then included, as per the United Kingdom's use of the term, the activities of individuals who influence the radicalization of others, and encourage others to engage in violent extremism (House of Commons, 2009-10; see Middle East Institute, 2015).

inclusion criteria were applied to both types of tweets. For details of the differences between retweets and directly authored tweets across the two samples, please refer to the Supplemental Materials.

The classifier used the linguistic style of those extremist Daesh tweets to calculate the probability that the 251,061 tweets in the remainder of the sample were written in this extremist (versus non-extremist) linguistic style. Once we had trained the learning algorithm on extremist tweets, we could proceed to the second step of our analysis.

The second analytic step was to predict the extent to which each individual user's tweets became more extremist over time and with the extent to which they engaged in mobilizing interactions. To do this, we applied the algorithm (Table 5) to the whole dataset (minus the training sample of 4,000 tweets; creating the test sample, n = 251,061). We saved the probability that each tweet was classified as extremist (due to how much it conformed to the linguistic style of the extremist tweets of Daesh supporters), and then used the logit of that probability (the logarithm of the odds) as a dependent variable in a subsequent multi-level linear mixed-effects model (LMM), in which tweets were nested within users. Through using this as a dependent variable in an LMM, we could predict whether an individual user changed in how much they conformed to the linguistic style of extremist tweets over time (e.g., became more similar to the tweets that expressed support for violent extremism, showing psychological change), and we could model the factors associated with those changes.

3.2 Results

3.2.1 Classifying Tweets Using Linguistic Profiles

Using the training subset of 4,000 tweets (2,000 human-rated extremist tweets from the Daesh sample and 2,000 randomly selected tweets from the baseline sample), we conducted a logistic regression to predict the class of tweet author (Daesh-supporter vs.

baseline user) from scores on the Daesh vernacular custom dictionary and function words dictionaries. The *null deviance* (*df* = 3999) of this training model was 5545.20; *residual deviance* (*df* = 3986) = 2413.70; *AIC* = 2441.70; *McFadden* R^2 = 0.56. To test the accuracy of the classifier, we conducted a 10-fold cross validation on the training set. Results of the 10-fold cross validation showed that the algorithm that included both function words and Daesh vernacular was 89% accurate on average (κ = 0.78; 95% *CI*s 0.88, 0.90), *F-score* = 0.89, *average precision* = 0.90 and *recall* = 0.88, and the area under the receiver operating characteristic curve (*AUROC*) was 0.94 (Table 5).

Table 5

Logistic Regression Results for Training set (*n* = 4,000).

Predictor	*B*	*SE*	*Odds ratio*	*Z*	*Wald*	*p*
Daesh vernacular	0.41	0.02	1.51	25.45	699.53	<0.001
Prepositions	0.11	0.01	1.11	12.62	159.29	<0.001
Third person plural pronouns	0.17	0.03	1.18	6.70	44.88	<0.001
Articles	0.08	0.01	1.08	6.21	38.53	<0.001
Auxiliary verbs	0.06	0.01	1.06	5.79	33.55	<0.001
Second person singular pronouns	-0.08	0.01	0.93	-5.30	28.10	<0.001
First person singular pronouns	-0.07	0.02	0.93	-4.28	18.33	<0.001
First person plural pronouns	-0.08	0.03	0.92	-3.07	9.41	0.002
Negations	-0.06	0.02	0.94	-3.04	9.22	0.002
Conjunctions	0.03	0.01	1.03	2.17	4.72	0.03
Impersonal pronouns	0.01	0.01	1.01	0.47	0.22	0.64
Quantifiers	0.01	0.02	1.01	0.36	0.13	0.72
Adverbs	-0.002	0.01	1.00	-0.20	0.04	0.84

Note. The reference group for interpreting the odds ratios is the Daesh supporter sample, i.e.,

for every one unit increase in use of Daesh vernacular, the odds of a tweet being classified as emanating from a Daesh supporter account increased by a factor of 1.51. Variables are listed in order of importance to the model (most: least).

Due to the volatility of vernacular over time relative to the stability of function words, we also tested whether the classifier could accurately predict the user account of tweets (Daesh vs. baseline) using an algorithm that only used the function words. The *residual deviance* (*df* = 3987) of this model was 3914.40; *AIC* = 3940.40; *McFadden* R^2 = 0.29. A 10-fold cross validation showed that the classifier was 79% accurate on average (κ = 0.58; *95% CIs* 0.78, 0.80); *F-score* = 0.78, *average precision* = 0.79 and *recall* = 0.78, *AUROC* = 0.85. A likelihood ratio test comparing the models showed that the algorithm that included Daesh vernacular fit the data better; $\chi^2(1) = 1500.80$, $p<.001$. However, taken together these statistics suggest that there is utility in using linguistic style to detect within-person changes online. Indeed, from an operational and ethical perspective, not having to be reliant on vernacular could be a useful feature.

However, because we were interested in predicting the extent to which each individual user's tweets became more extremist in terms of both *how* (function words) and *what* (Daesh vernacular) they wrote, we applied the algorithm that included function words and Daesh vernacular (Table 5) to the whole dataset (minus the training sample of 4,000 tweets; creating the test sample; 251,061 of tweets). We saved the probability that each tweet was classified as emanating from an Daesh-supporter account (due to how much it conformed to the linguistic style of the extremist tweets of Daesh supporters), and then used the logit of that probability (the logarithm of the odds) as a dependent variable in the subsequent LMMs.

3.2.2 Linear Mixed Model Predicting Within-Person Changes in Linguistic Style

To establish whether the users in the Daesh-supporter sample showed change over time relative to baseline, we conducted an LMM with grand mean centering and random slopes in

which the observations (251,061 tweets; Level 1) were nested within people (209 Twitter users; Level 2). The LMM assessed how the logit of the probability of individual users being classified as a Daesh supporter due to conformity to the linguistic style of extremist tweets changed over time (absolute time, indexed as account age in days and standardized across samples, and with chronological tweet number and the total volume of tweets sent by the user, both representing relative time), controlling for the users' number of sequential accounts (as a proxy for number of times the user had accounts suspended for violating Twitter's community standards). The inclusion of these variables enabled us to assess the impact of time in the presence of account suspensions. We also included a variable to distinguish between directly authored tweets and retweets, and the users' number of followers and the number of accounts that they followed, as proxies for their social network size. Results are available in Table 6. In LMM, absolute values of the *t* statistic greater than or equal to 1.96 indicate an effect that was significant at $p < .05$.

There was a main effect for sample, whereby Daesh-supporters were significantly more extremist than baseline. We found a significant interaction between sample (Daesh-supporter versus baseline) and account age (Figure 1), showing that conformity to the extremist linguistic style increased over time for Daesh-supporters ($\gamma = 0.02$, $p<.05$) but decreased for baseline users ($\gamma = -0.17$, $p<.05$). We also found a significant two-way interaction between sample and tweet type (directly authored versus retweet), whereby Daesh-supporters' directly authored tweets conformed significantly more to the extremist linguistic style than their retweets, $\gamma = 0.47$, $p<.05$, but baseline users' directly authored tweets conformed significantly less to the extremist linguistic style than their retweets, $\gamma = -0.79$, $p<.05$.

We then repeated this analysis using the algorithm that relied only on function words and not Daesh vernacular (that is, the LMM predicted changes in non-conscious linguistic

style only) and found the same results interaction between sample and tweet type, $\gamma = 0.22$, $p < .05$; but no interaction between sample and account age, $\gamma = -0.001$, $p > .05$.

Table 6

LMM Results Predicting Logit of the Probability of Daesh Class Membership (Based on Conformity with Extremist Linguistic Style) of Daesh Supporters and Baseline Users (Users, *N* = 209, Tweets, *n* = 251,061).

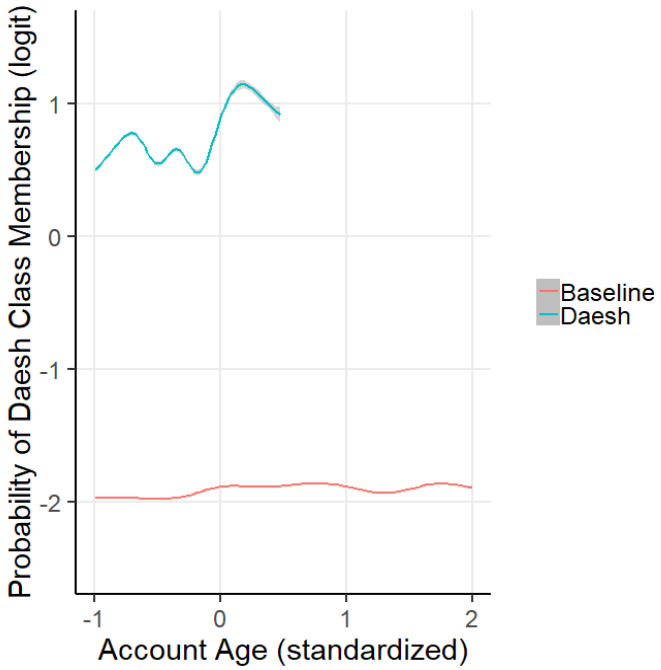
Predictor	*γ*	*SE*	*t*	*F*
Sample: Daesh-supporting (v. baseline)	2.05*	0.16	13.13	377.66
Tweet type: Directly authored (vs. retweeted)	0.47*	0.01	50.86	3964.69
Account age (days, standardized)	0.02*	0.01	2.39	15.12
Total tweet volume (standardized)	-0.26*	0.07	-3.78	15.31
Chronological tweet number (log. transformed)	-0.003	0.01	-0.52	0.20
Number of sequential accounts	0.25	0.17	1.47	2.49
Number of followers (standardized)	-0.01	0.10	-0.10	0.02
Number of accounts the user followed (standardized)	-0.03	0.04	-0.80	0.27
Sample x Tweet type	0.31*	0.02	14.17	200.78

Sample x Account age	0.15*	0.03	5.47	34.45

Note. *$p<.05$.

Figure 1

LMM Predicting Changes in the Probability of Individual Users being Classified as a Daesh Supporter over time.



Note. The line for the Daesh sample is shorter than the line for baseline accounts because Daesh-supporting accounts were younger than baseline accounts on average, and age of account was standardized across (rather than within) samples. This was necessary for comparability across samples over time in the LMM.

To investigate the mechanisms underlying the increases in conformity to the extremist linguistic style for Daesh-supporter accounts, we repeated the LMM using the Daesh-supporter data only (Table 7). Time (account age, in days, standardized within Daesh-supporting sample; controlling for relative time) was positively related to conformity with the extremist linguistic style, providing evidence for change over time (Figure 1). Mobilizing

interactions were also positively related to conformity to the extremist linguistic style[10], suggesting that interactions of this nature may provide a mechanism to explain these changes (Figure 2). Directly authored tweets were more extremist than retweets (see Supplemental Materials).

Table 7

LMM Results Predicting Logit of the Probability of Daesh Class Membership of Daesh Supporters (Users, *N* = 110, Tweets, *n* = 38,053).

Predictor	*γ*	*SE*	*t*	*F*
Tweet type: Directly authored (vs. retweet)	0.69*/0.35*	0.03/0.02	22.83/20.20	692.54/481.84
Account age (days, standardized)	0.14*/-0.02	0.05/0.03	2.47/-0.50	16.55/-0.53
Tweet volume (standardized)	-0.08/0.04	0.09/0.07	-0.93/0.51	-0.85/0.67
Mobilizing Interactions	0.53*/0.16*	0.01/0.01	39.93/20.94	1594.43/439.41
Chronological tweet number (log. transformed)	0.01/-0.01	0.02/0.01	0.30/-1.31	0.32/0.62
Number of sequential accounts	0.26/0.04	0.14/0.11	1.88/0.41	3.14/0.08
Number of followers (standardized)	1.97/2.23*	1.22/0.86	1.61/2.59	2.61/6.69
Number of accounts the user followed (standardized)	-0.15/-0.20*	0.11/0.08	-1.27/-2.54	0.09/0.42

Note. *p<.05. The number before '/' indicates the result for the algorithm that used both

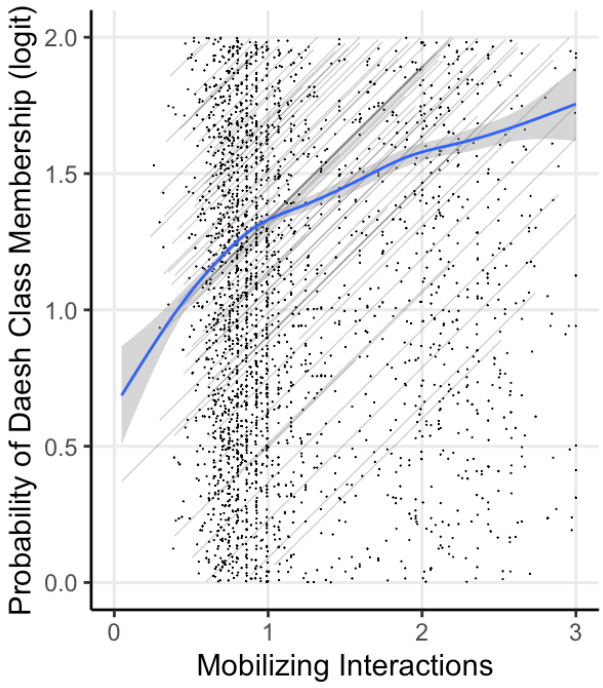
[10] To further verify that mobilizing interactions operates as a single construct, we ran an LMM in which we entered the dictionary categories anger, harm, collective action, threat, and twitter interactions separately alongside the other predictors (instead of the mobilizing interactions construct), predicting conformity with the extremist linguistic style. All positively predicted conformity, following the same pattern as the mobilizing interactions construct. Twitter interactions (@mentions) were negatively related to extremist classification. This may be an anomaly given that they positively correlate with harm and anger, and which are both positively related to conformity (Table 3). This unexpected result should be explored in future research.

Daesh vernacular and function words; the number after '/' indicates the result for the algorithm that used only function words.

We then repeated this analysis using the algorithm that relied only on function words and not Daesh vernacular and found the same results for mobilizing interactions and tweet type (Table 7). We also found a positive main effect for number of followers, and a negative main effect for number of accounts followed.

Figure 2

LMM Predicting Changes in the Probability of Individual Users being Classified as a Daesh Supporter via Mobilizing interactions.



Note. Three levels of data are indicated as follows: tweets (individual points), users (black lines), and sample (blue line). Mobilizing interactions represent the factor scores of a latent variable comprising talk of anger, threat, harm, collective action, and Twitter interactions. The shaded area around the target line represents the standard error.

To test for causal mediation, whereby conformity with the extremist linguistic style

increased over time (account age) through increases in mobilizing interactions (controlling for tweet type, relative time, number of sequential accounts, number of followers and number of accounts followed), we first fit the models for the mediator and outcome and then ran a causal mediation analysis (Imai, et al., 2010) in the Mediation package in *R* with 1000 simulations. The average causal mediation effect of account age on increases in conformity to the extremist linguistic style through mobilizing interactions was non-significant (γ = .02; 95% Quasi-Bayesian Confidence Intervals or *CI*s = -.005, .04; p = .14). The average direct effect of account age on conformity to the extremist linguistic style remained significant in the presence of the mediator (γ = .17; 95% *CI*s = .06, .28; p =.002); total effect = (γ = .19; 95% *CI*s = .08, .31; p <.001).

4. GENERAL DISCUSSION

Taken together, this research presents evidence that within-person changes in conformity to a group's vernacular and linguistic style can be captured in social media data. Across two studies, we present evidence that collective action intentions, and changes in both group vernacular and patterns of function word use, are positively associated with engaging in mobilizing interactions. As it is difficult for people to control their function word use in particular, this combination of findings supports our assertion that genuine psychological changes can occur through engaging in social media activity. Furthermore, for the first time, we demonstrated here that increases in conformity to extremist language can occur over time on social media platforms and are related to mobilizing interactions. These results therefore lend empirical weight both to politicians' assertions that psychological changes, such as radicalization, might be observable through social media activity (Department for Digital, Culture, Media, & Sport, 2019; UK Home Affairs Committee, 2017), and for the psychological role of the internet in terrorism (Gill et al., 2017; Taylor et al., 2017) – although this would be need to be verified in future research with ground truth. Ultimately,

by investigating social media interactions over time in this way, we can increase our understanding both of the power of social media interactions as a socialization mechanism, and also of the process by which individuals could become socialized into extremist psychological groups online. This is important, because if you know the mechanism for changes such as increases in endorsement of violence, you can design interventions to disrupt those changes.

Whilst Daesh supporter accounts varied in the extent to which they conformed to the vernacular and linguistic style of the gold-standard extremist tweets as a function of the extent to which they engaged in mobilizing interactions, this did not reflect a consistent and linear increase over time through those interactions. The direction of the effect is unclear: while past research suggests that mobilizing interactions may be a mechanism for psychological change, through increasing the extent to which people identify with a social change group or social movement (McGarty et al., 2014; Smith, Gavin, et al., 2015; Smith et al., 2018; Smith, Thomas, et al., 2015; Thomas et al., 2014; van Zomeren et al., 2008), it is also likely that people who identify with a group and for whom the group identity is salient (and show this through conforming more to the linguistic style) may engage in more mobilizing interactions as a result of conformity to ingroup norms (Tajfel & Turner, 1979; Turner, 1991). This is therefore likely to be a reciprocal process, and the absence of a direct linear relationship between mobilizing interactions and account age could reflect both the fluidity and dynamism of social identity salience over time – and the associated conformity to group norms/linguistic style.

Both Daesh vernacular plus function words, and function words alone, could be accurately used to predict the class from which the tweet emanated. However, while we observed changes in linguistic style over time when using the algorithm that included Daesh vernacular and function words, this was not observable when the algorithm relied on function

words only. This suggests that linguistic style is more stable over time than vernacular, and thus using a combination of function words and vernacular is more useful in assessing change.

We found that neither the Daesh-supporters' number of followers nor the number of users that they followed was significantly related to linguistic conformity when using the algorithm that included Daesh vernacular, but when using only function words the users' number of followers was positively related to linguistic conformity and the number of accounts they followed was negatively related to linguistic conformity. Overall, it did not appear that social isolation was (straightforwardly, at least) related to linguistic conformity.

4.1 Practical Implications

Our results are an adjunct to the proposal that radicalization trajectories can be detected using graph pattern matching algorithms to analyze fused data from social media and government/law enforcement agencies (Hung, et al., 2016; 2017), and research that attempts to identify signals for detecting terror threats (Brynielsson, Horndahl, Johansson, Kaati, Mårtenson, & Svenson, 2013). Importantly, our work differs in focus by elucidating the mechanisms of social media interactions and communication that predict within-person change, and thus could further enhance the discriminatory ability of such algorithms.

Now we have provided proof-of-concept for our method, it could be deployed to examine different types of changes. For example, future research should aim to distinguish between the mechanisms by which, and nature of, social media interactions that socialize individuals into violent extremist groups, and into (non-extremist) groups more broadly. This future research would require investigation into how interlocutors' perceptions of their intra- and intergroup context shape the content of their interactions, and thus the meaning of their shared psychological group membership and the kinds of group-normative actions they advocate.

The ability to model group socialization more broadly could bring insights to research and practice in other disciplines. For example, work in other domains has explored the identifying linguistic features associated with psychopathology (Junghaenel, Smyth, & Santner, 2008) such as schizophrenia (Bedi, et al., 2015; Lee, Lee, Ahn, & Kim, 2007). While the method we developed here (involving mobilizing interactions as a group socialization mechanism) is specific to modeling group socialization and detecting changes in psychological group memberships, it is possible that through social media interactions, people can develop social identification with others who are experiencing the same mental disorder (Haslam, Jetten, Cruwys, Dingle, & Haslam, 2018). According to Haslam and colleagues, this social bond with others who are experiencing similar mental health difficulties can provide a "social cure". Therefore, future research could explore the potential of the current method to be adapted for clinician use.

Most importantly, our evidence supports the notion that not only psychological state (e.g., Bedi, Carrillo, et al., 2015; Chung & Pennebaker, 2007; Losada, Crestani, & Parapar, 2017; Pennebaker, 2011), but changes in psychological state are detectable through natural language processing. Past research has been able to track general trends in word-use/semantics on social media platforms over time (for example, Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015), but this has not been done at the individual user level. Whilst Shrestha et al. (2017), Hung et al. (2017), and Abbasi and Chen (2005) showed that extremist messages can be detected using machine learning algorithms that discriminate using feature sets, again that past work did not index *changes* in those features in longitudinal data from individuals. In theory, in any context in which researchers or practitioners are interested in quantifying changes in psychological state, this method could be further explored. Validation with ground truth data would need to be the first step in this regard.

4.2 Limitations and ethical considerations

Our data and our conclusions are limited by the absence of ground truth data in this study (i.e., we do not have offline variables to triangulate with online behaviors). In fact, there are no open source ground truth datasets with which to validate a set of features associated with online support for Daesh (i.e., that includes the social media data of people known to be affiliated with Daesh offline – for example, people who have travelled to Syria to join Daesh or to become foreign fighters). Therefore, all research that claims to identify online posts by (and online features of) supporters of Daesh share this limitation. Without longitudinal ground truth data containing demographic variables and standardized psychological measures, we cannot investigate potential confounds, nor ascertain whether the changes in linguistic style are associated with other psychological changes. Furthermore, we cannot and do not wish to make any claims or assumptions about the users in this sample beyond the fact that each user explicitly and publicly expressed support for Daesh in their tweets (to the extent that the accounts were suspended for violating Twitter's terms of use). Future research should aim to explore how changes in these expressions relate to changes in psychological variables, such as social identification with the group and adherence to group norms. Given the absence of ground truth data, we have chosen not to publicly disclose the feature set or the Daesh vernacular dictionary due to the potential that they could be used to falsely identify individuals as supporters of violent extremism.

A second reason not to disclose the Daesh feature set is that there is no complete consensus on features across different research projects. Feature sets vary because of the lack of open source ground truth data with which to validate the features. Features are also time sensitive, so even if ground truth data were available, the features would become outdated as social media behavior evolved. Thus, the accuracy of the features and vernacular used in this study is likely to change over time. To establish the temporal stability of the features and their predictive validity, future research should use longitudinal ground truth data that

provides independent verification of how changes in group affiliation and offline behavior of the users relates to changes in features. Notwithstanding these limitations, the purpose of this research was to develop a new method to explore changes in linguistic style over time, rather than to create a time-independent, validated algorithm, feature set, or lexicon.

We should also be clear that the nature of our social media data necessarily limits the generalizability of our algorithm and feature set. Our Twitter data were comprised of short sentences (no longer than 140 characters) and were in English. The Twitter timelines spanned a relatively short time-frame and we used a relatively small number of accounts in our analyses. Therefore, the coefficients in our algorithm are specific to data with these features and cannot be generalized to different social media platforms, to the population of Daesh supporters more generally, or other forms of textual or spoken communication. Having said that, we were able to use this dataset to provide and train a relatively accurate classifier (even though accuracy *per se* was not our methodological goal), thus demonstrating that even with limited data, this method is viable.

Prior to data collection, Twitter suspended accounts for violating their terms of use. Whilst accounts had in common expression of support for Daesh, without ground truth we could not ascertain whether or not they were official Daesh media officers' accounts or unofficial supporter accounts. The changes in linguistic style we evidenced here therefore could have captured an evolving group narrative or style, or we could have captured changes that users made in response to Twitter's suspension activity. The latter is unlikely because users (as advised in Dabiq, Daesh' official publication) tended to use less, rather than more (as evidenced here) extremist features if they aimed to avoid suspension. Furthermore, whilst this might be the case for changes in vernacular, it is unlikely for linguistic style, which is more difficult to cultivate. More importantly however, notwithstanding this ambiguity about the causes of the changes we witnessed in our dataset, the key contribution of this research is

that we developed and tested a new paradigm with which such changes can be detected and modelled. That is, irrespective of the meaning behind the change, we provide here the first method that can capture such change. This paves the way for future research to isolate and validate linguistic patterns associated with psychological, social, and behavioral changes, respectively.

4.3 Conclusions

We contribute a significant advancement by presenting a method that can potentially model *changes in* expressions of extremism rather than detecting extremist posts *per se*. In doing so, we have created and validated a methodology and analytic strategy for extracting longitudinal psychological processes and evidence of within-person change from large samples of social media data. We have demonstrated that a relatively straightforward linguistic analysis can derive the feature vectors for Daesh-supporter classification, and that a linear function can be used to transform lexicon scores and social media metrics into a meaningful psychological factor capturing mobilizing interactions. Ultimately, this method can be adapted to any research context in which it is useful to capture within-person psychological change.

Author Contributions

First author: Conceptualization, Methodology, Investigation, Software, Formal analysis, Data curation, Writing-Original draft preparation, Supervision, Funding acquisition, Project administration.

Second author: Resources, Investigation, Software, Data curation, Formal analysis, Writing-Review & Editing.

Third author: Conceptualization, Methodology, Software, Writing-Review & Editing.

Fourth author: Conceptualization, Methodology, Writing-Review & Editing.

Fifth author: Investigation, Writing-Review & Editing.

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. *Ieee Intelligent Systems, 20*(5), 67-75. doi:10.1109/MIS.2005.81

Alimi, E. Y. (2006). Contextualizing political terrorism: A collective action perspective for understanding the Tanzim. *Studies in Conflict & Terrorism, 29*(3), 263-283. doi:10.1080/10576100600564216

Ashcroft, M., Fisher, A., Kaati, L., Omer, E., & Prucha, N. (2015). *Detecting Jihadist Messages on Twitter.* Los Alamitos: Ieee Computer Soc.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., . . . Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia, 1*, 15030. doi:10.1038/npjschz.2015.30

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Bott, L., Frisson, S., & Murphy, G. L. (2009). Interpreting conjunctions. *Quarterly Journal of Experimental Psychology, 62*(4), 681–706. https://doi.org/10.1080/17470210802214866

Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society, 3*(2), 1-12.

Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C., & Svenson, P. (2013). Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics, 2* (11), 1-15.

Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies, 41*, 230-233. doi:http://dx.doi.org/10.1016/j.electstud.2015.11.017

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science, 14*, 60-65.

Chen, H. C. (2008). *Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet.* New York: Ieee.

Chung, C. K., & Pennebaker, J. W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343-359). New York: Psychology Press.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educ Psychol Meas, 20*, 37–46.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, *WWW '11*, *745–754*, New York, NY, USA. ACM.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J. & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, 307–318*, New York, NY, USA, 2013. ACM.

Department for Digital, Culture, Media, & Sport. (2019). *Online Harms White Paper*. London: HM Government.

De Smedt, T., De Pauw, G., Van Ostaeyen, P. (2018). Automatic detection of online Jihadist hate speech. *CLiPS Technical Report Series 7, 1-31*, arXiv:1803.04596 [cs.CL]

Doosje, B., Moghaddam, F. M., Kruglanski, A. W., de Wolf, A., Mann, L., & Feddes, A. R. (2016). Terrorism, radicalization and de-radicalization. *Current Opinion in Psychology, 11,* 79-84. https://doi.org/10.1016/j.copsyc.2016.06.008

Europol. (2017). *EU Terrorism Situation and Trend Report (TE-SAT)*. Retrieved from https://www.europol.europa.eu/newsroom/news/2017-eu-terrorism-report-142-failed-foiled-and-completed-attacks-1002-arrests-and-142-victims-died

Executive Office of the President of the United States. (2016). *Strategic Implementation Plan for Empowering Local Partners to Prevent Violent Extremism in the United States*. Washington, DC: Executive Office of the President of the United States.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117–140.

Giles, H., Coupland, J., & Coupland, N. (1991). *Accommodation theory: Communication, context, and consequence*. Cambridge University Press, Cambridge.

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist use of the internet by the numbers. *Criminology & Public Policy*, n/a-n/a. doi:10.1111/1745-9133.12249

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research, 37*(1), 3-19. doi:10.1177/0093650209351468

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029-1046. doi:10.1037/a0015141

Grant, T., & Macleod, N. (2016). Assuming identities online: experimental linguistics applied to the policing of online paedophile activity. *Applied Linguistics, 37*(1), 50-70. doi:10.1093/applin/amv079

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998-1002. doi: 10.1126/science.1137651

Haslam, S. A. (2004). *Psychology in organisations: The social identity approach*. (2nd ed.). London: Sage.

Haslam, C., Jetten, J., Cruwys, T., Dingle, G., & Haslam, S. A. (2018). *The new psychology of health: unlocking the social cure*. London: Routledge.

Hou, W. K., & Hall, B. J. (2019). The mental health impact of the pro-democracy movement in Hong Kong. *Lancet Psychiatry, 6*(12), 982.

House of Commons and Local Government Committee. (2009-10). *Preventing Violent Extremism: Sixth report of session 2009-10*, 1,http://www.publications.parliament.uk/pa/cm200910/cmselect/cmcomloc/65/….

Hung, B. W. K., Jayasumana, A. P., & Bandara, V. W. (2016). *Detecting Radicalization Trajectories Using Graph Pattern Matching Algorithms*. New York: Ieee.

Hung, B. W. K., Jayasumana, A. P., & Bandara, V. W. (2017). *INSiGHT: A System for Detecting Radicalization Trajectories in Large Heterogeneous Graphs*. New York: Ieee.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309-334. doi:10.1037/a0020761

Interpol. (2016). *Annual Report 2016*. Retrieved from https://www.interpol.int/News-and-media/Publications2/Annual-reports2

Junghaenel, D. U., Smyth, J. M., & Santner, L. (2008). Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social and Clinical Psychology, 27*(1), 36-55. doi:10.1521/jscp.2008.27.1.36

Kaati, L., Omer, E., Prucha, N., & Shrestha, A. (2015*). Detecting multipliers of Jihadism on Twitter.* 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 954-960. New York: IEEE.

Khalil, J. (2017). The three pathways (3P) model of violent extremism. *The RUSI Journal, 162*(4), 40-48. doi:10.1080/03071847.2017.1365463

Khemlani, S., Orenes, I. & Johnson-Laird, P N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology, 24*, 5, 541-559. DOI:10.1080/20445911.2012.660913

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. *Paper presented at the Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, pp. 625-635. doi: 10.1145/2736277.2741627

Landis, J.R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1),159–174.

Lea, M., Spears, R., & Watt, S.E. (2007). Visibility and anonymity effects on attraction and group cohesiveness. *European Journal of Social Psychology, 37,* 761-773.

Lee, C. H., Lee, M., Ahn, S., & Kim, K. (2007). Preliminary analysis of language styles in a sample of schizophrenics. *Psychological Reports, 101*(2), 392-394. doi:10.2466/pr0.101.2.392-394

Lewin, K. (1953). Studies in group decision. In D. Cartwright & A. Zander (Eds.), *Group dynamics: Research and theory.* (pp. 287-301). New York: Row, Peterson & Co.

Losada, D. E., Crestani, F. & Parapar, J. (2017). eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In G.J.F. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 346–60. Springer International Publishing.

Lucic, L., & Bridges, E. (2018). Ecological landscape in narrative thought How siege survivors employ prepositions to make sense of war-torn Sarajevo. *Narrative Inquiry, 28*(2), 346-372. doi:10.1075/ni.17076.luc

Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior, 87*, 207-211. doi:https://doi.org/10.1016/j.chb.2018.05.035

Moskalenko, S. & McCauley, C. (2009). Measuring political mobilization: The distinction between activism and radicalism. *Terrorism and Political Violence, 21*(2), 239-260, DOI: 10.1080/09546550902765508

McCauley, C. & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence, 20*, 415-433. DOI: 10.1080/09546550802073367

McGarty, C., Thomas, E. F., Lala, G., Smith, L. G. E., & Bliuc, A. (2014). New technologies, new identities, and the growth of mass opposition in the Arab Spring. *Political Psychology, 35*(6), 725-740. doi:10.1111/pops.12060

Middle East Institute. (2015). *Deradicalization programs and counterterrorism: A perspective on the challenges and benefits*. Retrieved on 15th November, 2018 from https://www.mei.edu/publications/deradicalization-programs-and-counterterrorism-perspective-challenges-and-benefits

Morgan, G. S., Wisneski, D.C., & Skitka, L. J. (2011). The expulsion from Disneyland: The social psychological impact of 9/11. *American Psychologist, 66*,6, 447-454.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*(2), 125-135.

Paul, G., John, H., & Paige, D. (2014). Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists. *Journal of Forensic Sciences, 59*(2), 425-435. doi:10.1111/1556-4029.12312

Pennebaker, J. W. (2011). *The Secret Life of Pronouns*. London: Bloomsbury Press.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC [Computer software]*. . Austin, TX: LIWC.net.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312.

Pennebaker, J. W. & Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*, 3, 271-282.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577.

Postmes, T., Spears, R., Sakhel, K., & de Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behaviour. *Personality and Social Psychology Bulletin, 27*, 1243-1254.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Reicher, S., & Levine, M. (1994). Deindividuation, power relations between groups and the expression of social identity - the effects of visibility to the outgroup. *British Journal of Social Psychology, 33,* 145-163.

Reicher, S., & Levine, M. (1994). On the consequences of deindividuation for the strategic communication of self-identifiability and the representation of social identity. *European Journal of Social Psychology, 24*(4), 511-524.

Reicher, S., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology, 6*, 161-198.

Sageman, M. (2004). *Understanding Terror Networks* Philadelphia: University of Pennsylvania Press.

Sageman, M. (2017). *Turning to political violence: the emergence of terrorism*. Philadelphia: University of Pennsylvania Press.

Sassenberg, K., & Boos, M. (2003). Attitude change in computer-mediated communication: Effects of anonymity and category norms. *Group Processes & Intergroup Relations, 6*(4), 405-422. doi:10.1177/13684302030064006

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science, 8*(4), 379–386. https://doi.org/10.1177/1948550617715068

Scholand, A. J., Tausczik, Y. R., & Pennebaker, J. W. (2010). *Social Language Network Analysis.* New York: Assoc Computing Machinery.

Scrivens, R., Davies, G., & Frank, R. (2018). Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. *Behavioral Sciences of Terrorism and Political Aggression, 10*(1), 39-59. doi:10.1080/19434472.2016.1276612

Sedgwick, M. (2010). The concept of radicalization as a source of confusion. *Terrorism and Political Violence, 22*, 479 – 494. DOI: 10.1080/09546553.2010.491009

Shrestha, A., Kaati, L., & Cohen, K. (2017). A machine learning approach towards detecting extreme adopters in digital communities. In A. M. Tjoa & R. R. Wagner (Eds.), *2017 28th International Workshop on Database and Expert Systems Applications* (pp. 1-5). New York: Ieee.

Smith, L. G. E., Blackwood, L., & Thomas, E. (2020). The need to re-focus on the group as the site of radicalization. *Perspectives on Psychological Science*. https://doi.org/10.1177%2F1745691619885870

Smith, L. G. E., Gavin, J., & Sharp, E. (2015). Social identity formation during the emergence of the occupy movement. *European Journal of Social Psychology, 45*(7), 818-832. doi:10.1002/ejsp.2150

Smith, L. G. E., McGarty, C., & Thomas, E. F. (2018). After Aylan Kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time. *Psychological Science29(4), 623-634.* https://doi.org/10.1177/0956797617741107

Smith, L. G. E., & Postmes, T. (2009). Intra-group interaction and the development of norms which promote inter-group hostility. *European Journal of Social Psychology, 39*(1), 130-144. doi:10.1002/ejsp.464

Smith, L. G. E., & Postmes, T. (2011). The power of talk: Developing discriminatory group norms through discussion. *British Journal of Social Psychology, 50*(2), 193-215. doi:10.1348/014466610x504805

Smith, L. G. E., Thomas, E. F., & McGarty, C. (2015). "We must be the change we want to see in the world": Integrating norms and identities through social interaction. *Political Psychology, 36*(5), 543-557. doi:10.1111/pops.12180

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In S. Worchel & W. G. Austin (Eds.), *The social psychology of intergroup relations*. Monterey, CA: Brooks/Cole.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24-54. doi:10.1177/0261927X09351676

Taylor, P. J., Holbrook, D., & Joinson, A. (2017). Same kind of different: Affordances, terrorism, and the Internet. *Criminology and Public Policy, 16*(1), 127-133. doi:DOI:10.1111/1745-9133.12285

Thomas, E. F., McGarty, C., & Louis, W. R. (2014). Social interaction and psychological pathways to political engagement and extremism. *European Journal of Social Psychology, 44*(1), 15-22. doi:10.1002/ejsp.1988

Turner, J. C. (1991). *Social influence*. Milton Keynes: Open University Press.

UK Home Affairs Committee. (2017). *Hate Crime: Abuse, Hate and Extremism Online*. Retrieved from House of Commons: http://www.parliament.uk/homeaffairscom

van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin, 134*(4), 504-535. doi:10.1037/0033-2909.134.4.504